B-119E
c.2

Statistics Canada Statistique Canada

File reference No.
N/réf.

MEMORANDUM – *NOTE DE SERVICE*

Date March 1, 1983

To – *À* D. Binder, Chief, Health, Education, Science & Culture Survey Methods

From – *De* D. Paton, IASMD

Subject – *Objet* SMALL AREA ESTIMATION OF AGRICULTURE DATA USING REGRESSION TECHNIQUES

This memorandum is in two sections: the first summarizes the results of applying various linear model based regression techniques to the estimation for small areas using FES and Census of Agriculture data; the second is a discussion of some considerations relating to the use of linear model based inference techniques with survey data.

Tables and figures will be found at the end of the text.

9–8000–45: 19-4-82

SECTION 1

Methods for small area estimation have been documented by several authors (e.g. Drew, Singh, and Choudhry 1982; Kish 1980; Purcell and Kish 1979). In general, the techniques use auxilliary information and a knowledge of some structure in the population (relating the auxilliary information with the variables of interest) to refine estimates for the small areas of interest. The structure is usually defined as an approximate functional relationship between variables included in the auxilliary information and the variable(s) of interest. Examples of sources of auxilliary information are censuses and administrative records.

This study investigates the use of a linear model to approximate the structure relating data from the 1980 FES (in particular: fertilizer use) to auxilliary variables from the 1976 Census of Agriculture. The small areas of interest are crop districts.

Direct Estimates at the Crop District Level

The FES is a stratified design with at least two replicate samples drawn from each stratum. The replicates may be used to estimate the variance of the estimated stratum total. The strata do not cross crop district boundaries.

Tables 1 and 2 display the direct estimates of fertilizer use, by crop district and by province, calculated in the following way:

$\hat{y}$ crop district or province $= \Sigma\ \hat{y}$ stratum , where the summation is over strata in the crop district or province

$\hat{y}$ stratum $= \frac{1}{n_r} \Sigma\ \hat{y}$ replicate , where the summation is over the replicates in the stratum, and n_r is the number of replicates in the stratum

$\hat{y}$ replicate $= \Sigma \frac{1}{p_i} y_i$, where the summation is over the sampled units in the replicate, and p_i is the probability that unit i is selected

$\text{var}(\hat{y}$ c.d. or prov.$) = \Sigma\ \text{var}\ (\hat{y}$ stratum$)$

$\text{var}(\hat{y}$ stratum$) = \frac{1}{n_r} * \frac{1}{n_{r-1}} * \Sigma\ (\hat{y}$ replicate $- \hat{y}$ stratum$)^2$.

The tables show that the direct estimators produce provincial estimates with low coefficients of variation (for Manitoba, Saskatchewan and Alberta; B.C. consists in this analysis of a single crop district) while the c.v.'s for estimates at the crop district level are much higher.

Linear Regression Estimates

A number of approaches to the use of linear regression are available, including the design based one (Cochrane 1977), the (superpopulation) model based one, and a range of intermediate ones that consider the data from both points of view. Nathan (1981) discusses an intermediate approach and cautions against the model and design extremes.

This study started at the model based extreme and worked towards the intermediate approach.

Whatever approach is used, a decision must be made as to the level of aggregation to fit regressions to. Ideally, the largest aggregation that preserves the same correlation structure or linear model for the subpopulations of interest will be used. If a larger aggregation is used the estimates for subpopulations will be biased; if a smaller aggregation is used, the estimates of correlations or parameters will not be based on all the available information. In this study, it has been assumed that one regression describes fertilizer use for all farms on the prairies. This is a strong, and most likely unreasonable, assumption. Some evidence will be presented later that it is unreasonable.

The first set of regression estimates is from conventional analysis of the classical model:

$$y_i = X_i \underset{\sim}{\beta} + e_i \text{ , where the } e_i \text{ are i.i.d. } N(0,\sigma^2) \text{ and}$$

$\underset{\sim}{X}_i$ is a vector of census variables,

y_i is fertilizer use.

A stepwise procedure was used to select a specific set of carriers from the available census variables. The parameters of the model were estimated by:

$$\hat{\beta} = (X_n' X_n)^{-1} X_n' Y_n$$

$\hat{\sigma}^2$ = mean square residual

$$\hat{cov}(\hat{\beta}) = (X_n' X_n)^{-1} \hat{\sigma}^2$$

Estimates are calculated via:

$\hat{Y} = \underset{\sim}{X}\hat{\beta}$, where $\underset{\sim}{X} = \Sigma\, \underset{\sim}{x}_i$, the summation over farms in the small area of interest.

The sample weights are not used in this procedure.

Table 3 and Figures 1-3 compare crop district estimates and c.v.'s from this regression with the direct estimates. (The variance estimates of the regression estimates are model based and do not use the replication of the survey design.)

It is clear from Table 3 and Figure that (conditional on the model being true) this regression approach produces estimates with lower c.v.'s than those from the direct method. However, if the model is not true, these estimates may be biased, even for the grand total.

Nathan (1981) proposes that the weighted regression estimator: $\hat{\beta}_w = (X_n' W_n X_n)^{-1} X_n' W_n Y_n$ (where W_n is the diagonal matrix of inverses of inclusion probabilities) be used in preference to the unweighted one. This is an intermediate approach, as $\hat{\beta}_w$ is a consistent estimator of the usual parameter of the design based approach: $B = (X_N' X_N)^{-1} X_N' Y_N$ (the population correlation vector) as well as an unbiased estimator of β in the regression model.

Under the model, the covariance of $\hat{\beta}_w$ is larger than that for $\hat{\beta}$, however $\hat{\beta}_w$ is more robust to model failure. It gives unbiased estimates of the grand total even if the model fails.

When using the weighted estimator in a model-based analysis it is not clear how one should estimate the error variance: σ^2. Both the mean square residual (when the weights are normalized to sum to the sample size) are unbiased estimators of σ^2. When they agree we could use either, but when they disagree which should we use? Disagreement indicates that there is a correlation between the residuals and the weights, which may suggest that there is a problem with the model.

With the fertilizer use data, the weighted mean square residual (22, 273, 209) and the mean square residual (57, 504, 239) differ by a factor of about 2.5 so the c.v.'s will differ by a factor of about 1.6 depending on which estimate is used.

Tables 4 and 5, and figures 4-6 compare the direct, unweighted regression, and weighted regression estimates. Even in the worst case, using the weighted regression and the unweighted mean square residual to estimate the error variance, the model based approach produces estimates with generally lower c.v.'s than the design based direct estimator. A possible bias introduced by fitting one regression to all provinces is indicated by these tables and figures as the weighted regression estimates seem to consistently under-estimate fertilizer use in Manitoba and Alberta (compared to the direct estimator) and consistently over-estimate if in Saskatchewan.

A possible continuation of this study would be to fit separate models at the provincial or smaller level. An exploratory ANOVA-type analysis indicates that separate regressions could reasonably be fit at the crop district level (Table 6).

Another possible continuation would be to use the methods described in Binder (1982) to find design-based covariance estimates for $\hat{\beta}_y$. To have a means of comparing the accuracy of these various techniques, a variable from a census could be chosen as the variable of interest and transposed to the corresponding FES records. This type of analysis could then be repeated and the resulting domain, unweighted regression, and weighted regression estimates compared to the "true" value.

References

Binder, D.A. (1982), On the variances of asymptotically normal estimators from complex surveys. J. of Survey Methodology, Statistics Canada, 7, 157-170.

Cochrane, W.G. (1977), Sampling Techniques. John Wiley and Sons, New York

Drew, J.D., M.P. Singh, and G.H. Choudhry (1982), Evaluation of small area estimation techniques for the Canadian Labour Force Survey. Presented at the Annual Meeting of the American Statistical Association, Cincinnati, August 1982. 32 pp.

Kish, L. (1980), Design and estimation for domains. The Statistician, 29, 209-222.

Nathan, G. (1981), Notes on inference based on data from complex sample designs. J. of Survey Methodology, Statistics Canada, 7, 109-129.

Purcell, N.J. and L. Kish (1979), Estimation for small domains. Biometrics, 35, 365-384.

TABLE 1: Direct Estimates of Fertilizer Use, by Province

Province	Estimated Fertilizer Use	Coefficient of Variation (%)
Manitoba	117,596,098	5.85
Saskatchewan.	138,704,822	5.20
Alberta	165,389,125	4.57
B.C.	3,886,883	48.73
TOTAL	425,576,928	

TABLE 2: Direct Estimates of Fertilizer Use, by Crop District

Province*	Crop District	Estimated Fertilizer Use	Coefficient of Variation (%)
6	601	10520252	19.35
6	602	10639316	18.34
6	603	14150693	19.43
6	604	5407575	15.74
6	605	4908163	22.94
6	606	6305825	16.94
6	607	19893092	16.82
6	608	25621520	11.44
6	609	9637029	23.61
6	610	84360	64.71
6	611	7579028	24.57
6	612	2849245	19.22
7	701	8236972	16.03
7	702	4274673	17.39
7	711	3766326	30.75
7	712	8031288	17.61
7	721	2568378	17.28
7	722	1271693	21.31
7	733	710802	33.72
7	734	3350983	28.06
7	741	781356	29.76
7	742	1828344	43.31
7	751	9468944	15.42
7	752	11913309	15.42
7	761	10479278	20.50
7	762	6466425	19.87
7	771	6946918	23.61
7	772	6429123	26.37
7	781	8703194	26.25
7	782	14980490	18.88
7	791	13282563	20.88
7	792	15213761	16.32
8	810	4551946	33.02
8	820	32508184	8.39
8	830	21138247	15.45
8	841	15550309	16.78
8	842	23798571	15.71
8	850	32119956	9.18
8	860	14808950	13.91
8	870	20912961	8.32
9	909	3886883	48.74

* 6: Manitoba
7: Saskatchewan
8: Alberta
9: British Columbia

TABLE 3: Direct and Regression (Unweighted) Estimates of Fertilizer Use, by Crop District

Prov	Crop District	Direct Estimator		Regression Estimator			Delta[2.]
		Estimate	c.v.	Estimate	c.v.[1.] 1	2	
6	601	10520252	19.35	13233654	2.55	3.77	-1.33
6	602	10639316	18.34	10993422	2.59	4.48	-0.18
6	603	14150693	19.43	9772578	3.01	5.20	1.59
6	604	5407575	15.74	3329556	4.74	9.92	2.44
6	605	4908163	22.94	4156411	4.17	8.10	0.67
6	606	6305825	16.94	5869859	5.64	9.01	0.41
6	607	19893092	16.82	17333912	2.43	3.57	0.76
6	608	25621520	11.44	21071360	2.39	3.39	1.55
6	609	9637029	23.61	6893128	5.71	8.61	1.21
6	610	84360	64.71	582244	25.06	49.01	-9.12
6	611	7579028	24.57	4475135	6.15	10.07	1.67
6	612	2849245	19.22	3406958	9.82	14.99	-1.02
7	701	8236972	16.03	9749889	2.99	5.08	-1.15
7	702	4274673	17.39	7820392	3.85	6.51	-4.77
7	711	3766326	30.75	6630674	4.11	6.85	-2.47
7	712	8031288	17.61	10424685	3.93	5.83	-1.69
7	721	2568378	17.28	8242795	4.74	7.02	-12.79
7	722	1271693	21.31	4738709	4.66	8.21	-12.79
7	733	710802	33.72	5784576	4.62	7.43	-21.17
7	734	3350983	28.06	9038671	4.42	6.48	-6.05
7	741	781356	29.76	3432050	4.82	9.33	-11.40
7	742	1828344	43.31	4525834	5.18	8.52	-3.41
7	751	9468944	15.42	13633100	4.65	6.38	-2.85
7	752	11913309	15.42	18087789	4.13	5.43	-3.36
7	761	10479278	20.50	15961404	3.25	4.60	-2.55
7	762	6466425	19.87	11990251	3.77	5.59	-4.30
7	771	6946918	23.61	10922914	3.11	4.72	-2.42
7	772	6429123	26.37	10081240	2.57	4.65	-2.15
7	781	8703194	26.25	15945512	3.16	4.29	-3.17
7	782	14980499	18.88	15542723	3.16	4.41	-0.20
7	791	13282563	20.88	18170475	3.61	4.84	-1.76
7	792	15213761	16.32	13824001	2.75	4.30	0.56
8	810	4551946	33.02	9674240	4.10	6.08	-3.41
8	820	32508184	8.39	34130675	2.35	3.00	-0.59
8	830	21133247	15.45	25771308	3.07	3.88	-1.42
8	841	15550309	16.78	15529389	2.36	3.86	0.01
8	842	23798571	15.71	23414840	3.38	4.39	0.10
8	850	32119956	9.18	31781459	5.00	5.70	0.11
8	860	14808950	13.91	17643939	6.70	7.94	-1.38
8	870	20912961	8.32	35164118	3.32	3.87	-8.19
9	909	3886883	48.74	4616087	4.97	8.52	-0.38

1. c.v. 1 is the c.v. of the expected value
   c.v. 2 is the c.v. of the predicted value

2. Delta is the difference between the direct and regression estimates, in units of std. dev. of the direct estimate.

TABLE 4: Direct, and Weighted and Unweighted Regression Estimates of Fertilizer Use, by Crop District

Prov	Crop District	Direct Estimator		Regression Estimators				Delta	
				Unweighted		Weighted			
		Estimate	c.v.	Estimate	c.v.1	Estimate	c.v.1	Unweighted	Weighted
6	601	10520252	19.35	13233654	2.55	10477407	4.46	-1.33	0.02
6	602	10639316	18.34	10993422	2.59	8670665	4.53	-0.18	1.01
6	603	14150693	19.43	9772578	3.01	7695259	5.22	1.59	2.35
6	604	5407575	15.74	3329556	4.74	2633254	8.25	2.44	3.26
6	605	4908163	22.94	4156411	4.17	3150553	7.62	0.67	1.56
6	606	6305825	16.94	5869859	5.64	4565303	10.04	0.41	1.63
6	607	19893092	16.82	17333912	2.43	13770976	4.23	0.76	1.83
6	608	25621520	11.44	21071360	2.39	16549786	4.20	1.55	3.10
6	609	9637029	23.61	6893128	5.71	5509981	9.96	1.21	1.81
6	610	84360	64.71	582244	25.06	507237	40.84	-9.12	-7.75
6	611	7579028	24.57	4475135	6.15	3613264	10.66	1.67	2.13
6	612	2849245	19.22	3406958	9.82	2525235	18.75	-1.02	0.59
7	701	8236972	16.03	9749889	2.99	7876869	5.12	-1.15	0.27
7	702	4274673	17.39	7820392	3.85	6270843	6.56	-4.77	-2.68
7	711	3766326	30.75	6630674	4.11	5486255	6.85	-2.47	-1.49
7	712	8031288	17.61	10424685	3.93	8548675	6.60	-1.69	-0.37
7	721	2568378	17.28	8242795	4.74	6658265	8.09	-12.79	-9.22
7	722	1271693	21.31	4738709	4.66	3877110	7.89	-12.79	-9.61
7	733	710802	33.72	5784576	4.62	4734728	7.81	-21.17	-16.79
7	734	3350983	28.06	9038671	4.42	7410592	7.44	-6.65	-4.32
7	741	781356	29.76	3432050	4.82	2645609	8.82	-11.40	-8.02
7	742	1828344	43.31	4525834	5.18	3589796	9.17	-3.41	-2.22
7	751	9468944	15.42	13633100	4.65	11009947	7.89	-2.85	-1.06
7	752	11913309	15.42	18087789	4.13	14211498	7.21	-3.36	-1.25
7	761	10479278	20.50	15961404	3.25	12898809	5.51	-2.55	-1.13
7	762	6466425	19.87	11990251	3.77	9534461	6.52	-4.30	-2.39
7	771	6946918	23.61	10922914	3.11	8898956	5.26	-2.42	-1.19
7	772	6429123	26.37	10081240	2.57	7984072	4.42	-2.15	-0.92
7	781	8703194	26.25	15945512	3.16	11889130	5.77	-3.17	-1.39
7	782	14980490	18.88	15542723	3.16	11619357	5.77	-0.20	1.19
7	791	13282563	20.88	18170475	3.61	13884511	6.48	-1.76	-0.22
7	792	15213704	16.32	13824001	2.75	10363190	4.99	0.56	1.95
8	810	4551946	33.02	9674240	4.10	7314158	7.74	-3.41	-1.84
8	820	32508184	8.39	34130675	2.35	27094541	4.13	-0.59	1.98
8	830	21138247	15.45	25771308	3.07	20420064	5.44	-1.42	0.22
8	841	15550309	16.78	15529889	2.36	11938515	4.19	0.01	1.38
8	842	23798571	15.71	23414840	3.38	18430975	5.88	0.10	1.44
8	850	32119956	9.18	31781459	5.00	24599086	9.11	0.11	2.55
8	860	14808950	13.91	17643939	6.70	13587759	12.22	-1.38	0.59
8	870	20912961	8.32	35164118	3.32	26769828	6.10	-8.19	-3.37
9	909	3886883	48.74	4616087	4.97	3631417	8.94	-0.38	0.13

TABLE 5: Direct, and Weighted and Unweighted Regression Estimates of Fertilizer Use, by Province

Prov	Direct Estimator		Regression Estimators			
	Estimate	c.v.	Unweighted Estimate	c.v.1	Weighted Estimate	c.v.1
MAN.	117,596,098	5.85	101,118,217	3.31	79,668,918	5.79
SASK.	138,704,822	5.20	214,547,683	3.31	169,392,673	5.71
ALTA.	165,389,125	4.57	193,110,467	3.28	150,154,927	5.83
B.C.	3,866,883	48.74	4,616,087	4.97	3,631,417	8.95
TOTAL	425,576,928	-	513,392,454	-	402,847,935	-

TABLE 6: Comparison of One Line Model to Two Many Line Models

Model		d.f.	Sum of Squares	
A: One Line	Regression	12	219×10^6	
	Residual	7409	416×10^6	$R^2 = .33$
B: Five Lines *1	Regression	62	298×10^6	
	Residual	4359	337×10^6	$R^2 = .47$
C: 41 Lines *2	Regression	175	364×10^6	
	Residual	7246	271×10^6	$R^2 = .57$

Comparing A to B:

$$\frac{\frac{79}{50}}{\frac{337}{7359}} = 34.5$$

$$F_{30,120.005} = 1.98$$

Comparing B to C:

$$\frac{\frac{66}{113}}{\frac{271}{7246}} = 15.6$$

$$F_{120,120.005} = 1.61$$

*1 : separate lines fit to each of five groups of crop districts

*2 : separate lines fit to each crop district

FIGURE 1: DIRECT VS REGRESSION ESTIMATES OF FERTILIZER USE, BY CROP DISTRICT

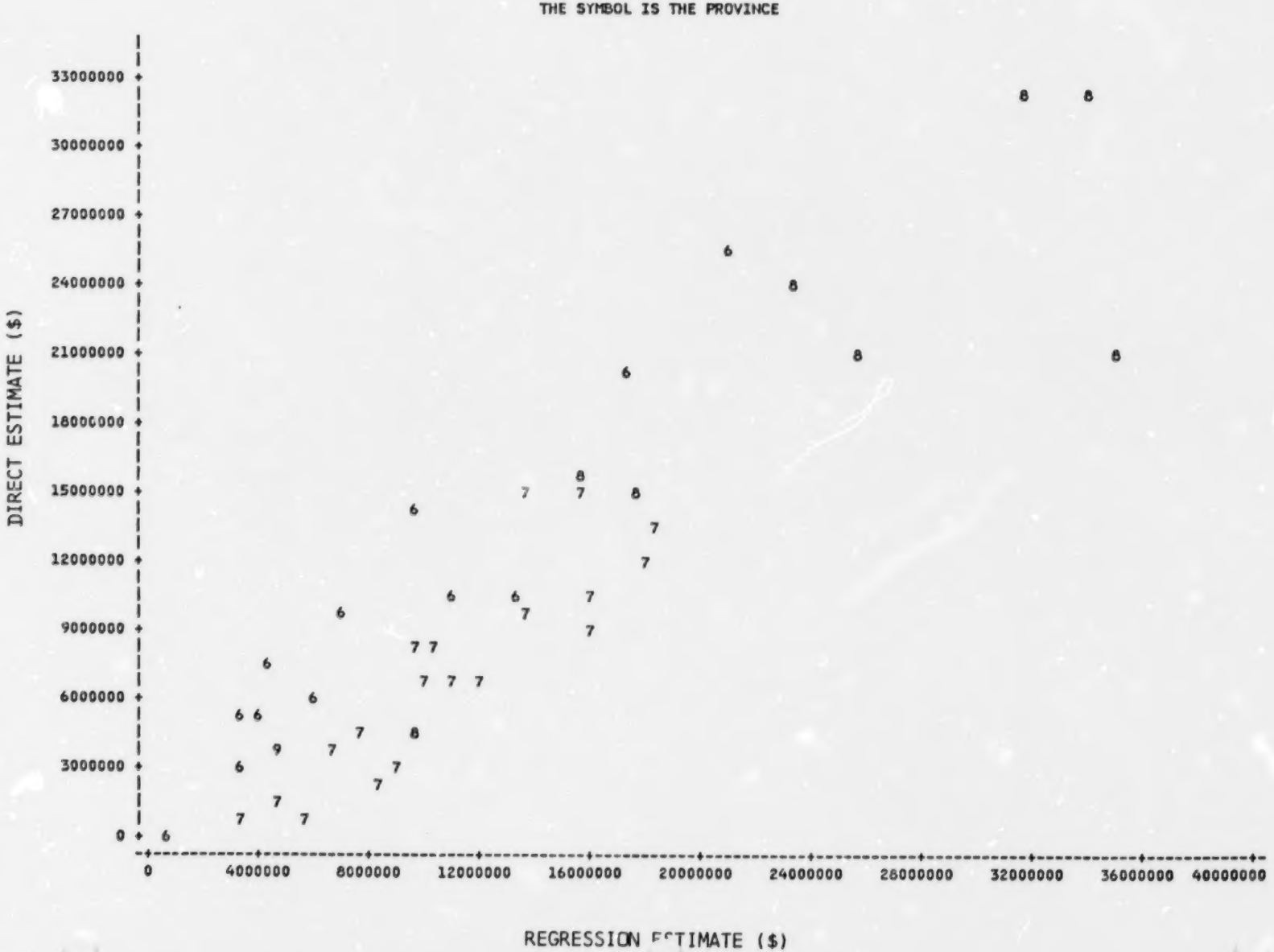


FIGURE 2: DIRECT AND REGRESSION ESTIMATES OF FERTILIZER USE, BY CROP DISTRICT

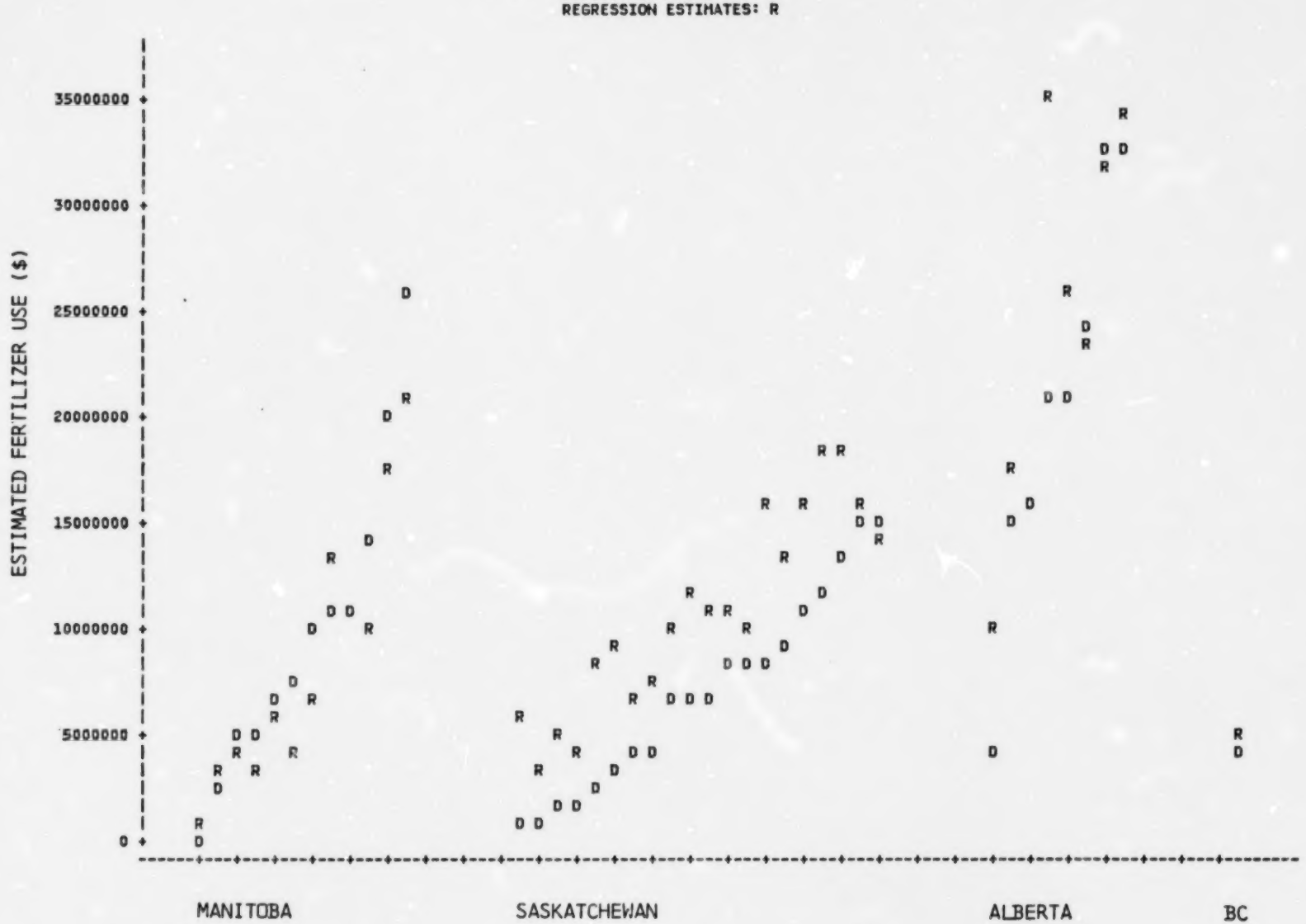


FIGURE 3: DIRECT AND REGRESSION ESTIMATE C.V.'S
BY CROP DISTRICT

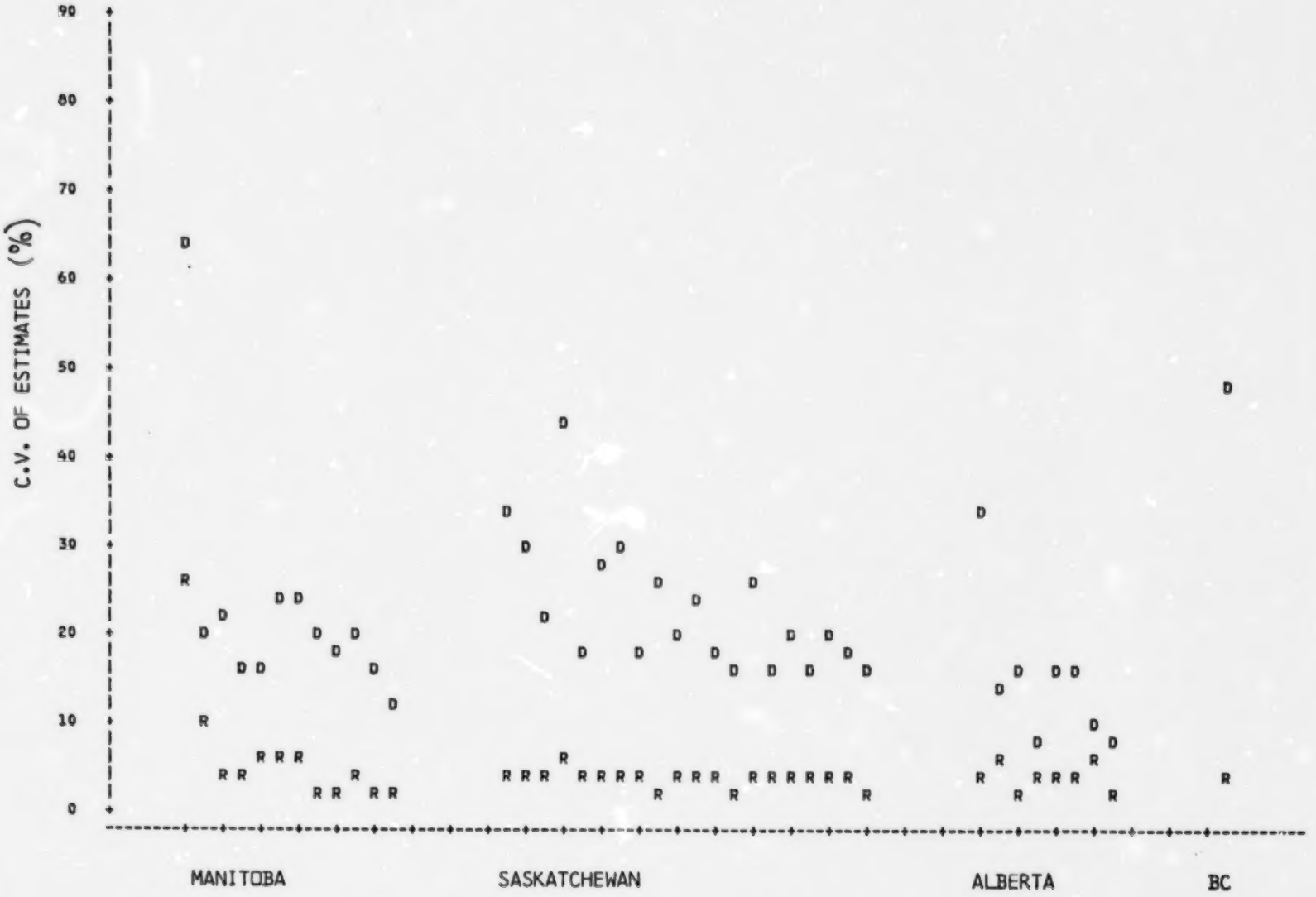


FIGURE 4: DIRECT VS WEIGHTED REGRESSION ESTIMATES OF FERTILIZER USE, BY CROP DISTRICT

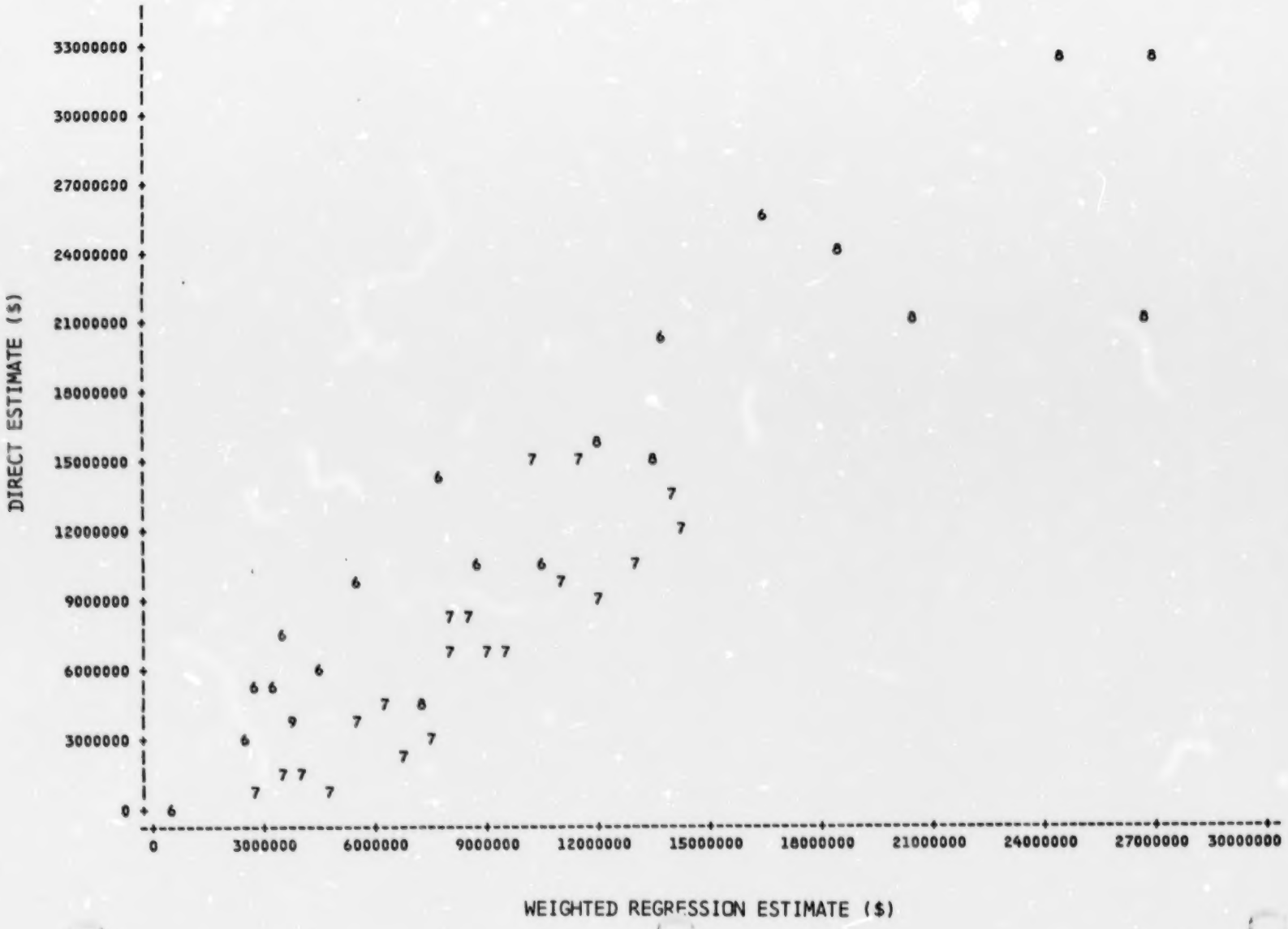


FIGURE 5: DIRECT, AND UNWEIGHTED AND WEIGHTED REGRESSION ESTIMATES OF FERTILIZER USE, BY CROP DISTRICT

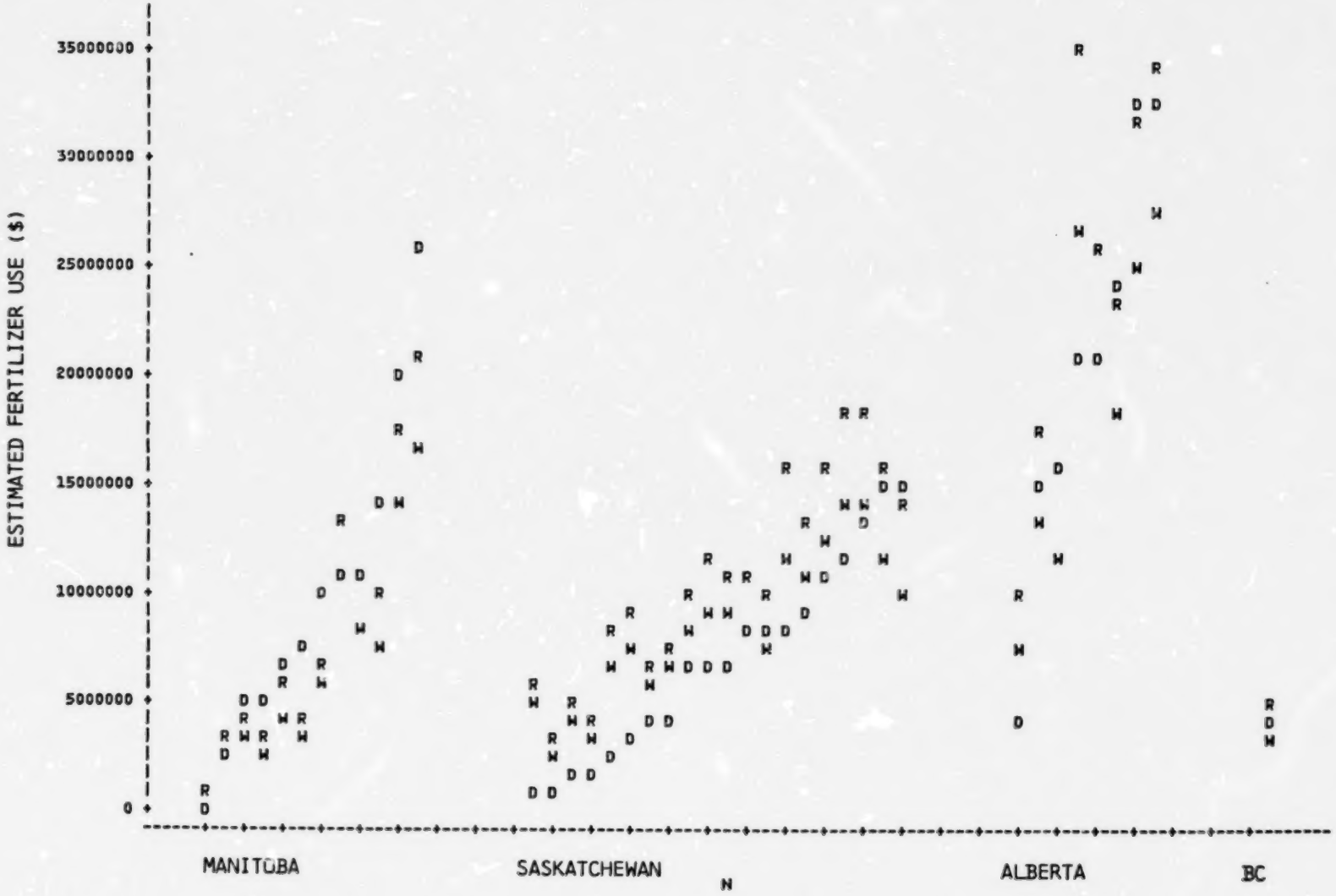
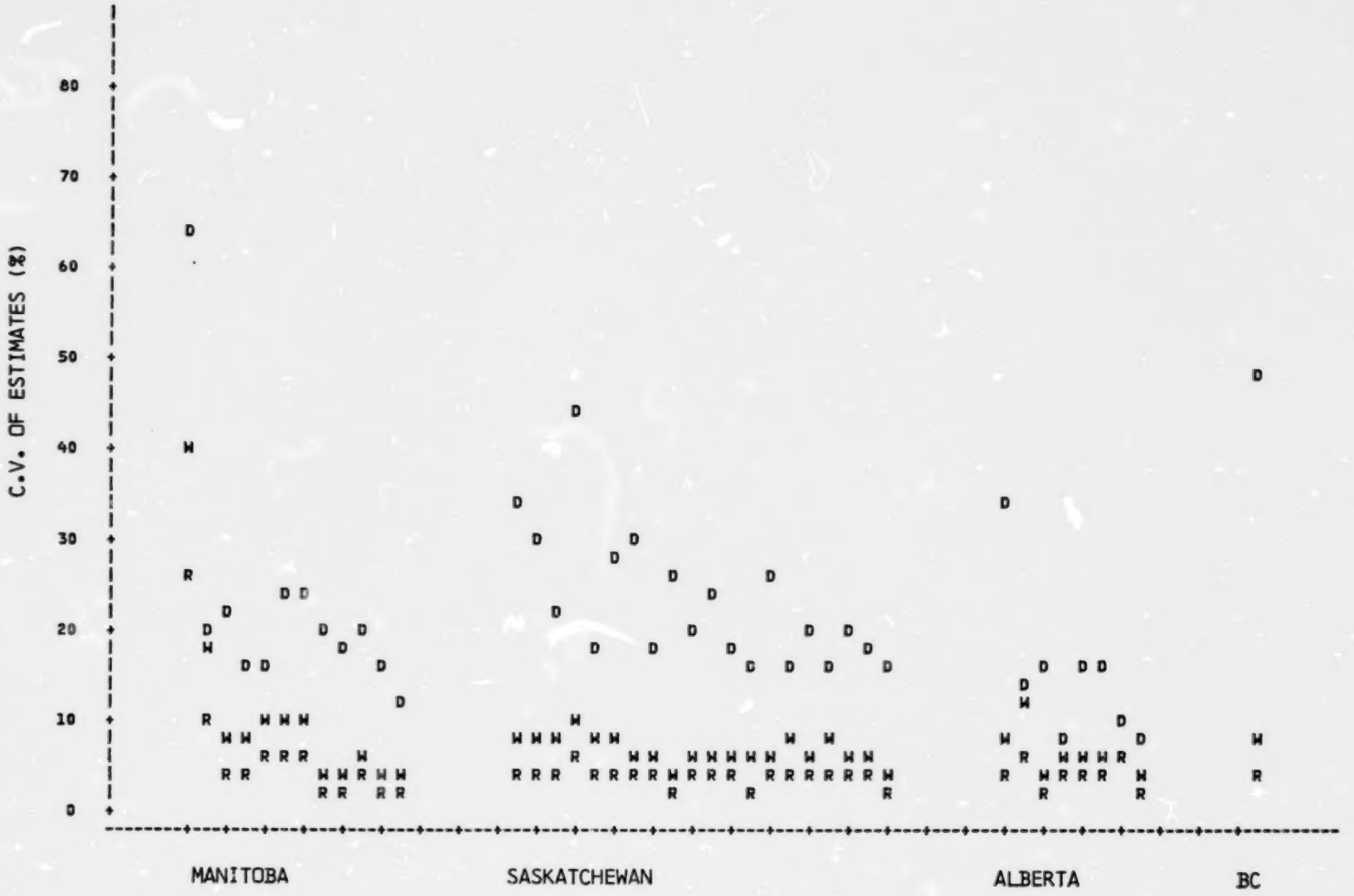


FIGURE 6: DIRECT, AND UNWEIGHTED AND WEIGHTED
REGRESSION ESTIMATE C.V.'S ,
BY CROP DISTRICT

DIRECT ESTIMATES: D
UNWEIGHTED REGRESSION ESTIMATES: R
WEIGHTED REGRESSION ESTIMATES: W



SECTION 2

The following is a rambling, informal discussion of some points relating to the use of linear models for inference from sample surveys.

In the introduction of Nathan (1981) the author states: "The first question which must be answered by anyone who intends to carry out statistical analysis is what exactly are the parameters about which inference is required". Exactly what parameters are of interest will depend on the specific model the statistician chooses or is asked to use.

In the traditional survey analysis approach the design based one, the population is seen as a set of fixed numbers, not random variables. A random component is introduced through the sample design.

The parameters of interest are finite population parameters, functions of the fixed population. I think of this point of view as being analogous to an accountants interest in the actual incomes and expenses of a business.

The superpopulation approach, in contrast, assumes that the population is a realization from a probability distribution. The parameters of interest may be related to this underlying distribution rather than the specific realization. I think of this point of view as being analogous to that of the managers of a business. They may be interested in modelling incomes and expenses in terms of a parametric probability distribution. They do not do this because they believe the model is true, in fact they believe it to be false, but rather to help them understand the processes that determine incomes and expenses.

Identifying exactly the parameters of interest may help determine whether the user is more interested in the realization or in the process that generated the realization.

Even when a design based approach is used, the finite population parameters often correspond to estimators of superpopulation parameters. It may make sense to estimate these finite population parameters, even when the population is not compatible with the corresponding superpopulation model; however, when this is done, one must be careful to interpret the results of the analysis as relating only to the population and not as providing insight into possible mechanisms that give us this population.

The range of possible superpopulation models is effectively unlimited. One of the most common statistical models is the linear model. The finite population parameters that correspond: ratios, correlations, population regression coefficients; are common subjects of design based inference. In the following discussion of linear models much of the notation follows that of Nathan (1981).

Let us call the variable of interest Y, and the vector of carrier variables X (the auxilliary variables). Our choice of a linear model relating X and Y means that we believe:

$$Y \doteq X' \beta$$

for some unknown vector, under some interpretation of $\doteq$.

If by ' $\doteq$ ' we mean:

$$Y_i = X_i' \beta + e_i$$

where the e_i are independent Gaussian variables with zero mean and variance known up to a constant multiplier (i.e. $e_i \sim N(0, \sigma^2 k_i)$, k_i known) we have the classical model for linear regression. In this case, inference follows standard paths and is relatively straightforward. Sampling weights would not be used in the analysis, and if k_i = constant, β would be estimated by:

$$\hat{\beta} = (X_n' X)^{-1} X_n' Y_n$$

If, however, by ' $\doteq$ ' we mean something somewhat different, the route to inference is not so obvious. Other meanings of $Y \doteq X'\beta$ might be:

a. $Y_i = X_i'\beta + e_i$ where the e_i are not dependent, Gaussian, etc.

b. $Y_i = f(X_i)$ with no random component, but where $|f(X_i) - X_i'\beta|$ is in some sence "small" on average.

c. $Y_i = f(X_i) + e_i$ the e_i being a random component, $|f(X_i) - X_i'\beta|$ "small".

The third model is quite general, and includes the other two as special cases, as well as the classical model. Our beliefs under this model can be summarized as:

(1) There is a functional relationship between the variable of interest and the carrier variables, perhaps with a random component added.

(2) This functional relationship can be approximated with a linear one.

(3) Our real interest is in the "true" functional relationship, but we are prepared to make inferences about a linear approximant as substitute.

The question of what parameters we wish to estimate in this case requires careful consideration, as well as how we decide that one set of parameters is better or worse than another. The "best" linear approximant of a function will very likely depend on the distribution of carrier values we are interested in. (For instance, if we want to approximate $f(X) = X^2$ with a line, the "best" line if we are only interested in positive X will be different from that if our interest is restricted to negative X.)

In a linear regression setting, the classical means of comparing two sets of parameters is by the sum of the squared residuals. This means of comparison is reasonable under the classifcal model with gaussian errors, but is not necessarily reasonable in other situations; it is, however, computationally attractive.

It will often be the case that the distribution of carriers over which we want to approximate f(X) will correspond with their distribution in the sampled population, and that for computational reasons as much as anything else we wish to use the sum of squared residuals to compare parameter estimates. In this case, the weighted estimator:

$$\hat{\beta}_w = (X_n' W_n X_n)^{-1} X_n' W_n Y_n$$

adjusts for over-representation in the sample of those units with high selection probabilities. If the classical model is true and $k_i = 1$ then the variance of $\hat{\beta}_w$ is

$$\sigma^2(X_n' W_n X_n)^{-1} X_n' W_n X_n (X_n' W_n X_n)^{-1} .$$

This is somewhat larger than the variance of $\hat{\beta}$, but note that under this classical model:

$$E(\hat{\beta}_w) = (X_n' W_n X_n)^{-1} X_n' W_n (X_n' \beta) = \beta$$

so $\hat{\beta}_w$ is still unbiased.

In addition $\hat{\beta}_w X_n$ is a sample design consistent estimator of Y_N while $\hat{\beta} X_N$ is not. Thus $\hat{\beta}_w$, the weighted estimator, seems to have a number of advantages over $\hat{\beta}$.

To make model based inference we still require an estimate of the error variance: σ^2. The expected value of the squared residuals includes a contribution from error and one from bias $(f(X) - X'\beta)$. If the bias is small relative to the error, it may be reasonable to ignore it, and to use the squared residuals to estimate σ^2. In this case should sampling weights be used in the estimation of σ^2? If they are to be used, they should be normalized somehow, perhaps to sum to the sample size (i.e. so that the average weight is one). The unweighted variance of the residuals (using the appropriate divisor) seems to be a reasonable estimate of σ^2 as well. (In fact comparison of the weighted and

unweighted variance estimates may be a more sensitive diagnostic tool for checking the adequacy of the linear, constant variance model than the difference between weighted and unweighted coefficient estimates suggested by Nathan (1981) since the fit may be essentially the same despite differences in the coefficients.)

All this discussion presumes that the carriers to be included in the model have been selected, and that a decision has been made as to what subset of the surveyed units will be used to fit this model. Neither of these decisions is a straight-forward one. Standard stepwise techniques can help indicate the model to use, but the F statistics and significance levels are incorrect if standard packages are used to perform weighted regressions. The subset of surveyed units to fit the model to must be selected by balancing the conflicting requirements of minimizing bias and of obtaining precise estimates of β by using all of the available information. Thus the subset will presumably include all the surveyed units from the domain of interest (or all of one class of surveyed units) but may additionally include surveyed units from other domains.

B-119F
c. 2.

Statistics Canada Statistique Canada

Memorandum Note de service

File reference no
Numero de reference

Date Le 1er mars 1983

To / À D. Binder, Chef, Section de la santé, de l'éducation, des sciences et de la culture

From / De D. Paton, Division des méthodes d'enquête-institutions et agriculture

Subject / Objet ESTIMATION DES DONNÉES SUR L'AGRICULTURE POUR LES PETITES RÉGIONS PAR DES TECHNIQUES DE RÉGRESSION

Cette note se divise en deux sections: la première présente un sommaire des résultats de l'application de diverses techniques de régression fondées sur un modèle linéaire à l'estimation pour les petites régions à partir de données de l'EEA et du recensement de l'agriculture; la deuxième porte sur certains aspects de l'utilisation de techniques d'inférence fondées sur un modèle linéaire pour analyser des données d'enquêtes.

Des tableaux et des figures sont présentés à la fin du texte.

SECTION 1

Plusieurs auteurs ont décrit des méthodes d'estimation pour les petites régions (par exemple, Drew, Singh et Choudhry 1982; Kish 1980; Purcell et Kish 1979). En général, ces techniques utilisent des information auxiliaires et des renseignements sur une certaine structure qui existe dans la population (structure qui précise le rapport entre les informations auxiliaires et les variables dépendantes) pour améliorer la qualité des estimations calculées pour les petites régions. Cette structure est habituellement définie sous la forme d'une relation fonctionnelle approximative entre certaines variables auxiliaires et la ou les variables dépendantes. A titre d'exemples de sources d'informations auxiliaires, mentionnons les recensements et les dossiers administratifs.

La présente étude examine l'utilisation d'un modèle linéaire pour produire une approximation de la relation structurelle entre des données de l'EEA de 1980 (en particulier, les données sur l'utilisation des engrais) et des variables auxiliaires tirées du recensement de l'agriculture de 1976. Les petites régions dont il est question ici sont les districts agricoles.

Estimations directes à l'échelle des districts agricoles

L'EEA est fondée sur un plan d'échantillonnage stratifié, et au moins deux échantillons répétés sont tirés de chaque strate. Ces échantillons répétés peuvent servir à l'estimation de la variance d'un total estimé pour une strate. Aucune strate ne dépasse les limites d'un district agricole.

Les tableaux 1 et 2 présentent les estimations directes de l'utilisation des engrais selon la province et le district agricole. Ces chiffres proviennent des calculs suivants:

$\hat{y}$ (district agricole ou province)	=	$\Sigma\hat{y}$ (strate)	où la sommation comprend chaque strate du district agricole ou de la province
$\hat{y}$ (strate)	=	$\frac{1}{n_r}\Sigma\hat{y}$ (échantillon répété)	où la sommation comprend chaque échantillon répété dans la strate et n_r est le nombre d'échantillons répétés dans la strate
$\hat{y}$ (échantillon répété)	=	$\Sigma \frac{1}{p_i} y_i$	où la sommation comprend chaque unité échantillonnée de l'échantillon répété et p_i est la probabilité de sélection de l'unité i
var($\hat{y}$ (district agricole ou province))	=	Σ var ($\hat{y}$ (strate))	
var($\hat{y}$ (strate))	=	$\frac{1}{n_r} * \frac{1}{n_r-1} * \Sigma$ ($\hat{y}$ (échantillon répété) - $\hat{y}$ (strate))^2.	

Les tableaux montrent que les estimateurs directs produisent des estimations provinciales qui ont un coefficient de variation faible (dans le cas du Manitoba, de la Saskatchewan et de l'Alberta; la Colombie-Britannique est représentée par seulement un district agricole dans cette analyse), tandis que les CV des estimations calculées à l'échelle des districts agricoles sont beaucoup plus élevés.

Estimations de régression linéaire

Il existe un bon nombre de méthodes de régression linéaire dont, entre autres, la régression basée sur un plan d'échantillonnage (Cochran 1977), la régression basée sur un modèle (de superpopulation) et une gamme de méthodes intermédiaires qui analysent les données en combinant ces points de vue. Nathan (1981) décrit une approche intermédiaire et souligne les dangers des extrêmes basés sur soit un modèle, soit un plan d'échantillonnage.

La présente étude a eu comme point de départ une méthode basée sur un modèle, mais elle s'est par la suite orientée vers une optique intermédiaire.

Indépendamment de la méthode utilisée, il faut décider du degré d'aggrégation pour l'adjustement des courbes de régression. Idéalement, on doit utiliser le plus haut degré d'agrégation qui reproduit la même structure de corrélation ou le même modèle linéaire au niveau des sous-populations étudiées. Si le degré d'agrégation est plus élevé, les estimations calculées pour les sous-populations seront biaisées; si le degré d'agrégation est plus fiable, les estimations des corrélations ou des paramètres ne seront pas basées sur tous les renseignements qu'on possède. Dans la présente étude, il a été supposé qu'une seule régression décrit l'utilisation des engrais pour toutes les exploitations agricoles des Prairies. Cette hypothèse est forte et sans doute inexacte. On explique plus loin pour quelle raison cette hypothèse n'est pas raisonnable.

Le premier jeu d'estimations de régression repose sur le modèle classique:

$y_i = X_i \underset{\sim}{\beta} + e_i$, où les résidus e_i sont indépendants et identiquement distribués selon une loi de type $N(0,\sigma^2)$, X_i est un vecteur de variables de recensement et y_i est la quantité d'engrais utilisée.

Un procédé pas à pas a été utilisé pour sélectionner un ensemble précis de variables exogènes à partir de l'éventail des données de recensement. Les paramètres du modèle ont été estimés par:

$$\hat{\beta} = (X_n' X_n)^{-1} X_n' Y_n$$

$\hat{\sigma}^2$ = la moyenne des carrés des résidus

$$\text{côv}(\hat{\beta}) = (X_n' X_n)^{-1} \hat{\sigma}^2$$

Les estimations sont calculées à partir de la formule:

$\hat{Y} = \underset{\sim}{X}\hat{\beta}$, où $\underset{\sim}{X} = \Sigma x_i$; la sommation comprend chaque ferme dans la petite région sous étude.

Les poids d'échantillonnage ne sont pas utilisés dans ce procédé.

Le tableau 3 et les figures 1 à 3 permettent de comparer les estimations pour les districts agricoles et les CV de cette régression avec les résultats de l'estimation directe. (Les estimations de la variance des estimations de régression sont basées sur le modèle et ne tiennent pas compte des échantillons répétés du plan d'échantillonnage.)

Il ressort clairement du tableau 3 et de la figure 2 que (à condition que le modèle soit vrai) cette méthode de régression produit des estimations qui ont des CV plus faibles que ceux des estimations directes. Toutefois, si le modèle n'est pas vrai, ces estimations peuvent être biaisées, même dans le cas du total global.

Nathan (1981) propose l'utilisation de l'estimateur de régression pondérée $\hat{\beta}_w = (X_n' W_n X_n)^{-1} X_n' W_n Y_n$ (où W_n est la matrice diagonale des inverses des probabilités d'inclusion) de préférence à l'estimateur de régression non pondérée. Cet estimateur relève d'une méthode intermédiaire, étant donné que $\hat{\beta}_w$ est une estimateur convergent du paramètre habituel de la technique basée sur le plan d'échantillonnage, $B = (X_N' X_N)^{-1} X_N' Y_N$ (le vecteur des corrélations de la population) et aussi de β dans le modèle de régression.

Dans le modèle de régression, la covariance de $\hat{\beta}_w$ est plus grande que celle de $\hat{\beta}$, mais $\hat{\beta}_w$ est plus robuste que $\hat{\beta}$ si le modèle n'est pas juste. Même dans une telle situation, $\hat{\beta}_w$ produit des estimations non biaisées du total global.

Lorsqu'on utilise l'estimateur pondéré dans une analyse basée sur un modèle, il n'est pas clair de quelle façon il faut estimer la variance résiduelle σ^2. La moyenne des carrés des résidus de la régression pondérée (quand les poids sont normalisés de façon que leur somme soit égale à la taille de l'échantillon) et celle de la régression non pondérée sont toutes les deux des estimateurs centrés de σ^2. Quand ces deux moyennes des carrés concordent, on peut utiliser l'une ou l'autre mais, dans le cas contraire, laquelle des deux faut-il choisir? Une différence entre ces moyennes des carrés indiquerait une corrélation entre les résidus et les poids, ce qui peut signifier une lacune dans le modèle.

En ce qui concerne les données sur l'utilisation des engrais, la différence entre la moyenne des carrés des résidus de la régression pondérée (22, 273, 209) et celle de la régression non pondérée (57, 504, 239) est un facteur de 2.5; la différence entre les CV sera donc un facteur d'à peu près 1.6, selon l'estimation qu'on choisit.

Les tableaux 4 et 5 et les figures 4 à 6 présentent une compariason des estimations directes, des estimations de la régression non pondérée et de celles de la régression pondérée. Même dans le pire scénario où on effectue une régression pondérée et calcule la moyenne des carrés des résidus de la régression non pondérée pour estimer la variance résiduelle, la méthode basée sur le modèle produit des estimation dont les CV sont généralement plus faibles que dans le cas de l'estimateur direct basé sur le plan d'échantillonnage. Ces tableaux et ces figures semblent indiquer

la possibilité d'un biais introduit par l'ajustement d'une courbe de régression pour l'ensemble des provinces. En effet, les estimations de la régression pondérée semblent systématiquement sous-évaluer l'utilisation des engrais au Manitoba et en Alberta (si on compare ces chiffres avec les estimations directes), de même qu'elles la surévaluent en Saskatchewan.

Une suite possible à la présente étude serait d'ajuster des modèles distincts à l'échelon provincial ou infraprovincial. Une vérification préliminaire basée sur l'analyse classique de la variance indique qu'il serait raisonnable d'ajuster des courbes de régressions individuelles au niveau des districts agricoles (tableau 6).

Une autre suite possible serait d'utiliser les méthodes décrites par Binder (1982) pour calculer des estimations de la covariance de $\hat{\beta}_w$ basées sur le plan d'échantillonnage. Pour pouvoir comparer l'exactitude de ces diverses techniques, une variable de recensement pourrait être choisie comme variable dépendante et transposée sur les enregistrements correspondants de l'EEA. Cette forme d'analyse pourrait alors être répétée, et chaque estimation calculées pour les domaines considérés par la régression non pondérée et la régression pondérée pourrait être comparés avec la "vraie" valeur.

Bibliographie

Binder, D.A. (1982), Les variances d'estimateurs asymptotiquement normaux basés sur des enquêtes complexes. Version française d'une article paru dans Techniques d'enquête 7, 157-170 (Statistique Canada).

Cochran , W.G. (1977), Sampling Techniques. John Wiley and Sons, New York.

Drew, J.D., M.P. Singh et G.H. Choudhry (1982), Evaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active au Canada. Document rédigé en prévision des réunions annuelles de l'American Statistical Association, Cincinnati, août 1982. 36 pages.

Kish, L. (1980), Design and estimation for domains. The Statistician 29, 209-222.

Nathan, G. (1981), L'inférence statistique basée sur des plans d'échantillonnage complexes. Version française d'un article paru dans Techniques d'enquête 7, 109-129 (Statistique Canada).

Purcell, N.J. et L. Kish (1979), Estimation for small domains. Biometrics 35, 365-384.

TABLEAU 1: Estimations directes de l'utilisation des engrais, par province

Province	Utilisation estimée des engrais	Coefficient de variation
		(%)
Manitoba	117,596,098	5.85
Saskatchewan.	138,704,822	5.20
Alberta	165,389,125	4.57
C.-B.	3,886,883	48.73
TOTAL	425,576,928	

TABLEAU 2: Estimations directes de l'utilisation des engrais, par district agricole

Province*	District agricole	Utilisation estimée des engrais	Coefficient de variation (%)
6	601	10520252	19.35
6	602	10639316	18.34
6	603	14150693	19.43
6	604	5407575	15.74
6	605	4908163	22.94
6	606	6305825	16.94
6	607	19893092	16.02
6	608	25621520	11.44
6	609	9637029	23.61
6	610	84360	64.71
6	611	7579028	24.57
6	612	2849245	19.22
7	701	8236972	16.03
7	702	4274673	17.39
7	711	3766326	30.75
7	712	8031288	17.61
7	721	2568378	17.28
7	722	1271693	21.31
7	733	710802	33.72
7	734	3350983	23.06
7	741	781356	29.76
7	742	1828344	43.31
7	751	9468944	15.42
7	752	11913309	15.42
7	761	10479278	20.50
7	762	6466425	19.87
7	771	6946918	23.61
7	772	6429123	26.37
7	781	8703194	26.25
7	782	14980490	18.68
7	791	13282563	20.68
7	792	15213761	16.32
8	810	4551946	33.02
8	820	32508184	8.39
8	830	21138247	15.45
8	841	15550309	16.78
8	842	23798571	15.71
8	850	32119956	9.18
8	860	14808950	13.91
8	870	20912961	8.32
9	909	3886883	48.74

* 6: Manitoba
7: Saskatchewan
8: Alberta
9: Colombie-Britannique

TABLEAU 3: Estimations directes et estimations de régression (non pondérée) de l'utilisation des engrais, par district agricole

Prov	District agricole	Estimateur direct		Estimateur de régresion			Delta[2]
		Estimation	CV	Estimation	CV[1] 1	CV[1] 2	
6	601	10520252	19.35	13233554	2.55	3.77	-1.33
6	602	10639316	18.34	10993422	2.59	4.48	-0.18
6	603	14150693	19.43	9772578	3.01	5.20	1.59
6	604	5407575	15.74	3329556	4.74	9.92	2.44
6	605	4908163	22.94	4156411	4.17	8.10	0.67
6	606	6305025	16.94	5869859	5.64	9.01	0.41
6	607	19393092	16.82	17333912	2.43	3.57	0.76
6	608	25621520	11.44	21071360	2.39	3.39	1.55
6	609	9637029	23.61	6893128	5.71	8.61	1.21
6	610	84360	64.71	582244	25.06	49.01	-9.12
6	611	7579028	24.57	4475135	6.15	10.07	1.67
6	612	2849245	19.22	3406958	9.82	14.99	-1.02
7	701	8236972	16.03	9749869	2.99	5.08	-1.15
7	702	4274673	17.39	7820392	3.85	6.51	-4.77
7	711	3766326	30.75	6630674	4.11	6.85	-2.47
7	712	8031288	17.61	10424685	3.93	5.83	-1.69
7	721	2568378	17.28	8242795	4.74	7.02	-12.79
7	722	1271693	21.31	4738709	4.66	8.21	-12.79
7	733	710302	33.72	5734576	4.62	7.43	-21.17
7	734	3350983	28.06	9038671	4.42	6.48	-6.05
7	741	781356	29.76	3432050	4.82	9.33	-11.40
7	742	1828344	43.31	4525834	5.18	8.52	-3.41
7	751	9468944	15.42	13633100	4.65	6.38	-2.85
7	752	11913309	15.42	18087789	4.13	5.43	-3.36
7	761	10479278	20.50	15961404	3.25	4.60	-2.55
7	762	6466425	19.87	11990251	3.77	5.59	-4.30
7	771	6946918	23.61	10922914	3.11	4.72	-2.42
7	772	6429123	26.37	10061240	2.57	4.65	-2.15
7	781	8703194	26.25	15945512	3.16	4.29	-3.17
7	782	14980490	18.88	15542723	3.16	4.41	-0.20
7	791	13282563	20.88	18170475	3.61	4.84	-1.76
7	792	15213761	16.32	13824001	2.75	4.30	0.56
8	810	4551946	33.02	9674240	4.10	6.08	-3.41
8	820	32508184	8.39	34130675	2.35	3.00	-0.59
8	830	21130247	15.45	25771303	3.07	3.88	-1.42
8	841	15550309	16.78	15529089	2.36	3.86	0.01
8	842	23766571	15.71	23414040	3.38	4.39	0.10
8	850	32119956	9.18	31781459	5.00	5.70	0.11
8	860	14803950	13.91	17643939	6.70	7.94	-1.38
8	870	28912961	8.32	35164118	3.32	3.87	-8.19
9	909	3886883	48.74	4616087	4.97	8.52	-0.38

1. CV 1 est le CV théorique
   CV 2 est le CV de la valeur estimés

2. Delta est l'écart entre l'estimation directe et l'estimation de régression exprimé en unités de l'écart-type de l'estimation directe.

TABLEAU 4: Estimations directes et estimations de régression non pondérée et de régression pondérée de l'utilisation des engrais, par district agricole

Prov	District agricole	Estimateur direct		Estimateurs de régression				Delta	
				non pondéré		pondéré			
		Estimation	CV	Estimation	CV 1	Estimation	CV 1	Régression non pondéré	Régression pondérée
6	601	10520252	19.35	13233654	2.55	10477407	4.46	-1.33	0.02
6	602	10639316	18.34	10993422	2.59	8670665	4.53	-0.18	1.01
6	603	14150693	19.43	9772578	3.01	7695259	5.22	1.59	2.35
6	604	5407575	15.74	3329556	4.74	2633254	8.25	2.44	3.26
6	605	4908163	22.94	4156411	4.17	3150553	7.62	0.67	1.56
6	606	6305825	16.94	5869859	5.64	4565303	10.04	0.41	1.63
6	607	19893092	16.82	17333912	2.43	13770976	4.23	0.76	1.83
6	608	25621520	11.44	21071360	2.39	16549786	4.20	1.55	3.10
6	609	9637029	23.61	6893128	5.71	5509981	9.96	1.21	1.81
6	610	84360	64.71	582244	25.06	507237	40.84	-9.12	-7.75
6	611	7579028	24.57	4475135	6.15	3613264	10.66	1.67	2.13
6	612	2849245	19.22	3406958	9.82	2525235	18.75	-1.02	0.59
7	701	8236972	16.03	9749889	2.99	7876869	5.12	-1.15	0.27
7	702	4274673	17.39	7820392	3.85	6270843	6.56	-4.77	-2.68
7	711	3766326	30.75	6630674	4.11	5486255	6.85	-2.47	-1.49
7	712	8031088	17.61	10424685	3.93	8548675	6.60	-1.69	-0.37
7	721	2568378	17.28	8242795	4.74	6656265	8.09	-12.79	-9.22
7	722	1271693	21.31	4738709	4.66	3877110	7.89	-12.79	-9.61
7	733	710302	33.72	5784576	4.62	4734728	7.81	-21.17	-16.79
7	734	3350983	28.06	9038671	4.42	7410592	7.44	-6.05	-4.32
7	741	781356	29.76	3432050	4.82	2645609	8.82	-11.40	-8.02
7	742	1828344	43.31	4525934	5.18	3589796	9.17	-3.41	-2.22
7	751	9468944	15.42	13633100	4.65	11009947	7.89	-2.85	-1.06
7	752	11913389	15.42	18087789	4.13	14211498	7.21	-3.36	-1.25
7	761	10479278	20.50	15961404	3.25	12898809	5.51	-2.55	-1.13
7	762	6466425	19.87	11990251	3.77	9534461	6.52	-4.30	-2.39
7	771	6946918	23.61	10922914	3.11	8898956	5.26	-2.42	-1.19
7	772	6429123	26.37	10081240	2.57	7984072	4.42	-2.15	-0.92
7	781	8703194	26.25	15945512	3.16	11589130	5.77	-3.17	-1.39
7	782	14980490	18.88	15542723	3.16	11619357	5.77	-0.20	1.19
7	791	13202563	20.88	18170475	3.61	13834511	6.48	-1.76	-0.22
7	792	15213761	16.32	13824001	2.75	10363190	4.99	0.56	1.95
8	810	4551946	33.02	9674240	4.10	7314158	7.74	-3.41	-1.84
8	820	32508184	8.39	34130675	2.35	27094541	4.13	-0.59	1.98
8	830	21138247	15.45	25771308	3.07	20420064	5.44	-1.42	0.22
8	841	15550309	16.73	15529589	2.36	11930515	4.19	0.01	1.35
8	842	23790571	15.71	23414840	3.38	18430975	5.88	0.10	1.44
8	850	32110956	9.18	31781459	5.00	24599086	9.11	0.11	2.55
8	860	14808950	13.91	17643939	6.70	13587759	12.22	-1.33	0.59
8	870	20912961	8.32	35164118	3.32	26769828	6.10	-8.19	-3.37
9	909	3386883	48.74	4616087	4.97	3631417	8.94	-0.38	0.13

TABLEAU 5: Estimations directes et estimations de régression non pondérée et de régression pondérée de l'utilisation des engrais, par province

Prov	Estimateur Direct		Estimateurs de régression			
			non pondérée		pondérée	
	Estimation	CV	Estimation	CV 1	Estimation	CV 1
MAN.	117,596,098	5.85	101,118,217	3.31	79,668,918	5.79
SASK.	138,704,822	5.20	214,547,683	3.31	169,392,673	5.71
Alb.	165,389,125	4.57	193,110,467	3.28	150,154,927	5.83
C.-B.	3,866,883	48.74	4,616,087	4.97	3,631,417	8.95
TOTAL	425,576,928	-	513,392,454	-	402,847,935	-

TABLEAU 6: Comparaison d'un modèle à une courbe et de deux modèles à courbes multiples

Modèle		degrés de liberté	sommes des carrés	
A: Une courbe	Régression	12	219×10^6	
	Résidus	7409	416×10^6	$R^2 = .33$
B: Cinq courbes *1	Régression	62	298×10^6	
	Résidus	7359	337×10^6	$R^2 = .47$
C: 41 courbes *2	Régression	175	364×10^6	
	Résidus	7246	271×10^6	$R^2 = .57$

Comparaison de A et B:

$$\frac{\frac{79}{50}}{\frac{337}{7359}} = 34.5$$

$$F_{30,120.005} = 1.98$$

Comparaison de B et C:

$$\frac{\frac{66}{113}}{\frac{271}{7246}} = 15.6$$

$$F_{120,120.005} = 1.61$$

*1: courbes ajustées individuellement pour cinq groupes de districts agricoles

*2: courbes ajustées individuellement pour chaque district agricole

FIGURE 1: COMPARAISON ENTRE LES ESTIMATIONS DIRECTES ET LES ESTIMATIONS DE RÉGRESSION DE L'UTILISATION DES ENGRAIS, PAR DISTRICT AGRICOLE

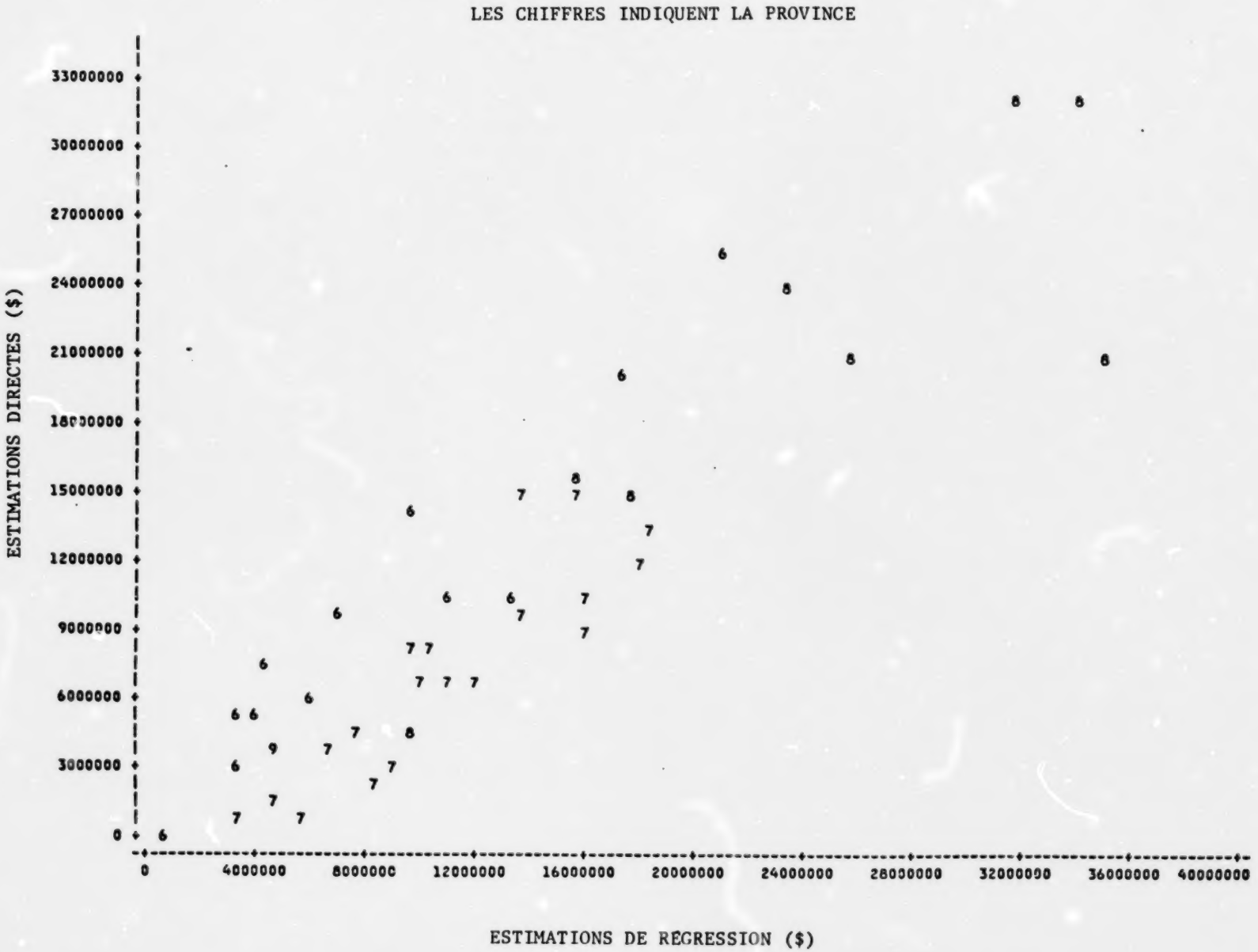


FIGURE 2: ESTIMATIONS DIRECTES ET ESTIMATION DE REGRESSION DE L'UTILISATION DES ENGRAIS, PAR DISTRICT AGRICOLE

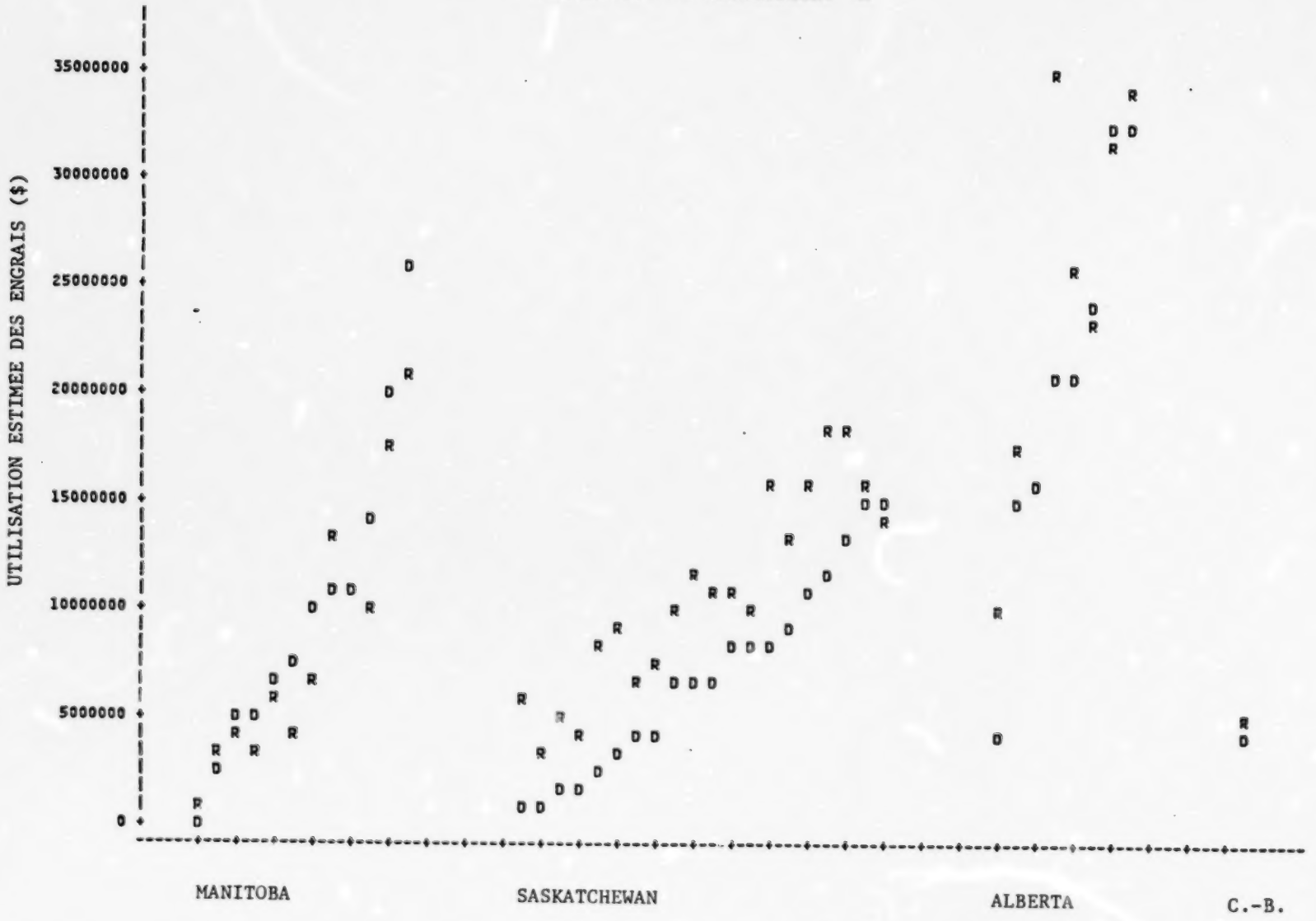


FIGURE 3: CV DES ESTIMATIONS DIRECTES ET DES ESTIMATIONS DE RÉGRESSIONS, PAR DISTRICT AGRICOLE

ESTIMATIONS DIRECTES: D
ESTIMATIONS DE RÉGRESSION: R

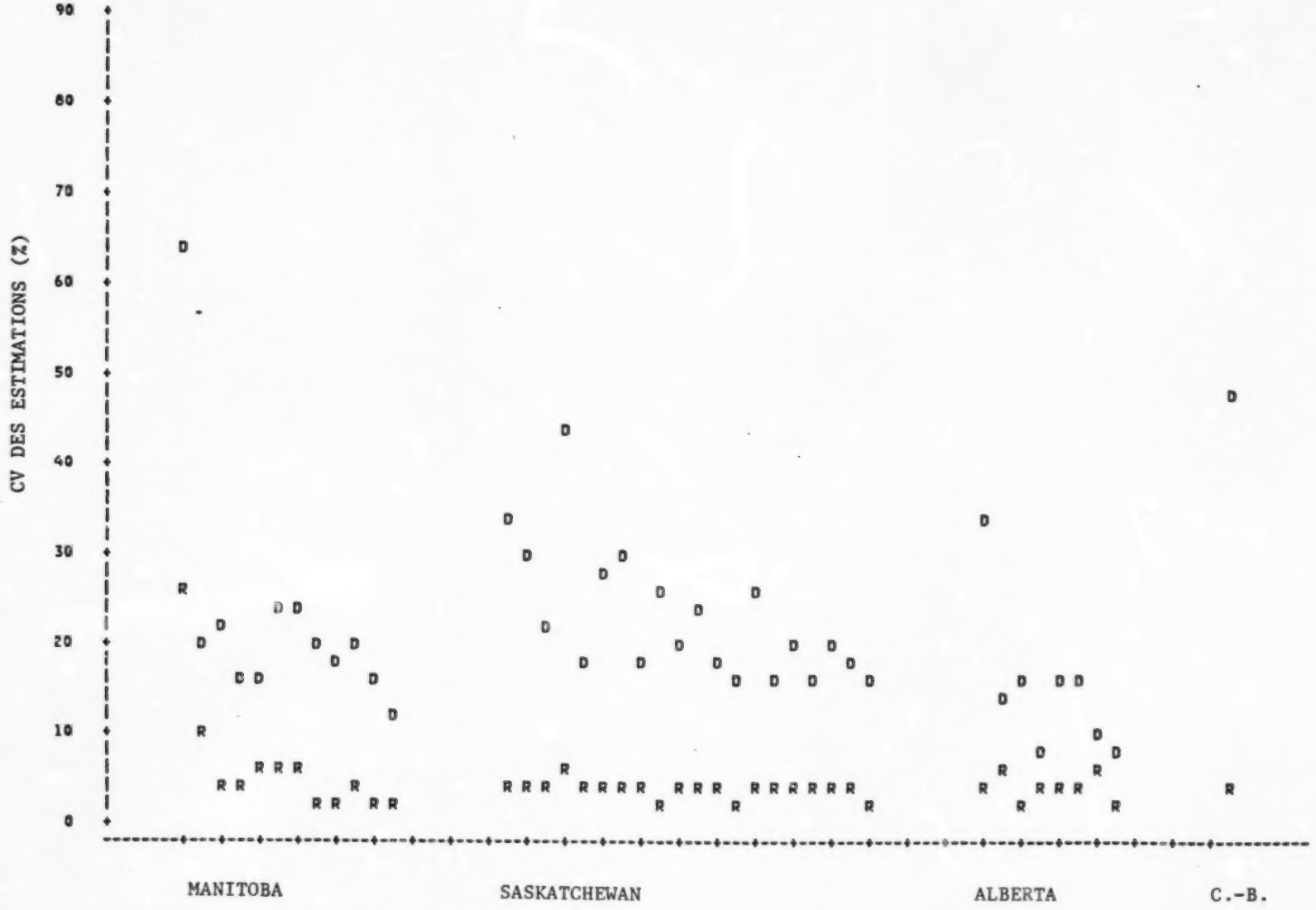


FIGURE 4: COMPARAISON DES ESTIMATIONS DIRECTES ET DES ESTIMATIONS DE REGRESSION PONDEREE DE L'UTILISATION DES ENGRAIS, PAR DISTRICT AGRICOLE

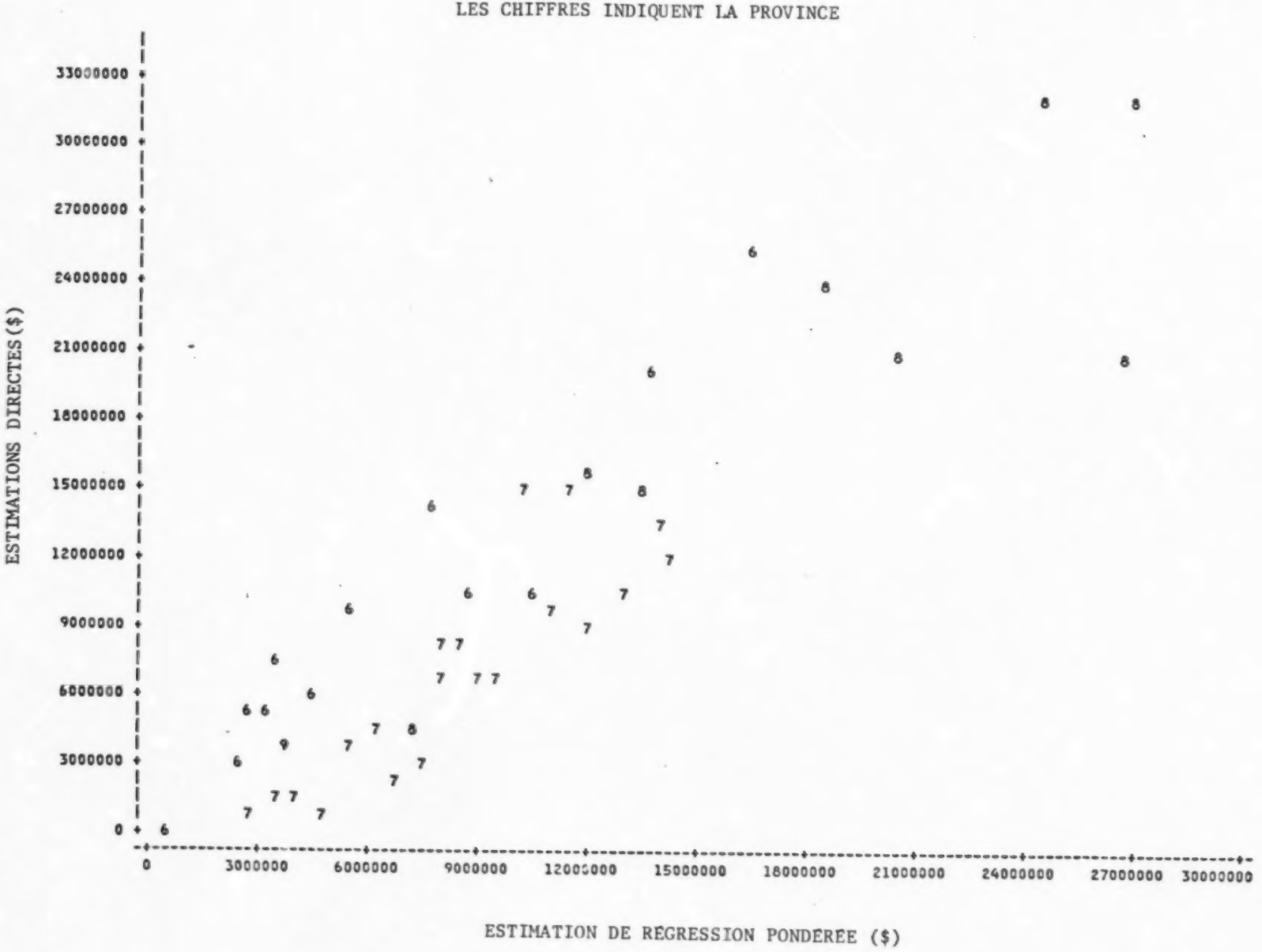


FIGURE 5: ESTIMATIONS DIRECTES, ESTIMATIONS DE RÉGRESSION NON PONDÉRÉE ET ESTIMATIONS DE RÉGRESSION PONDÉRÉE DE L'UTILISATION DES ENGRAIS, PAR DISTRICT AGRICOLE

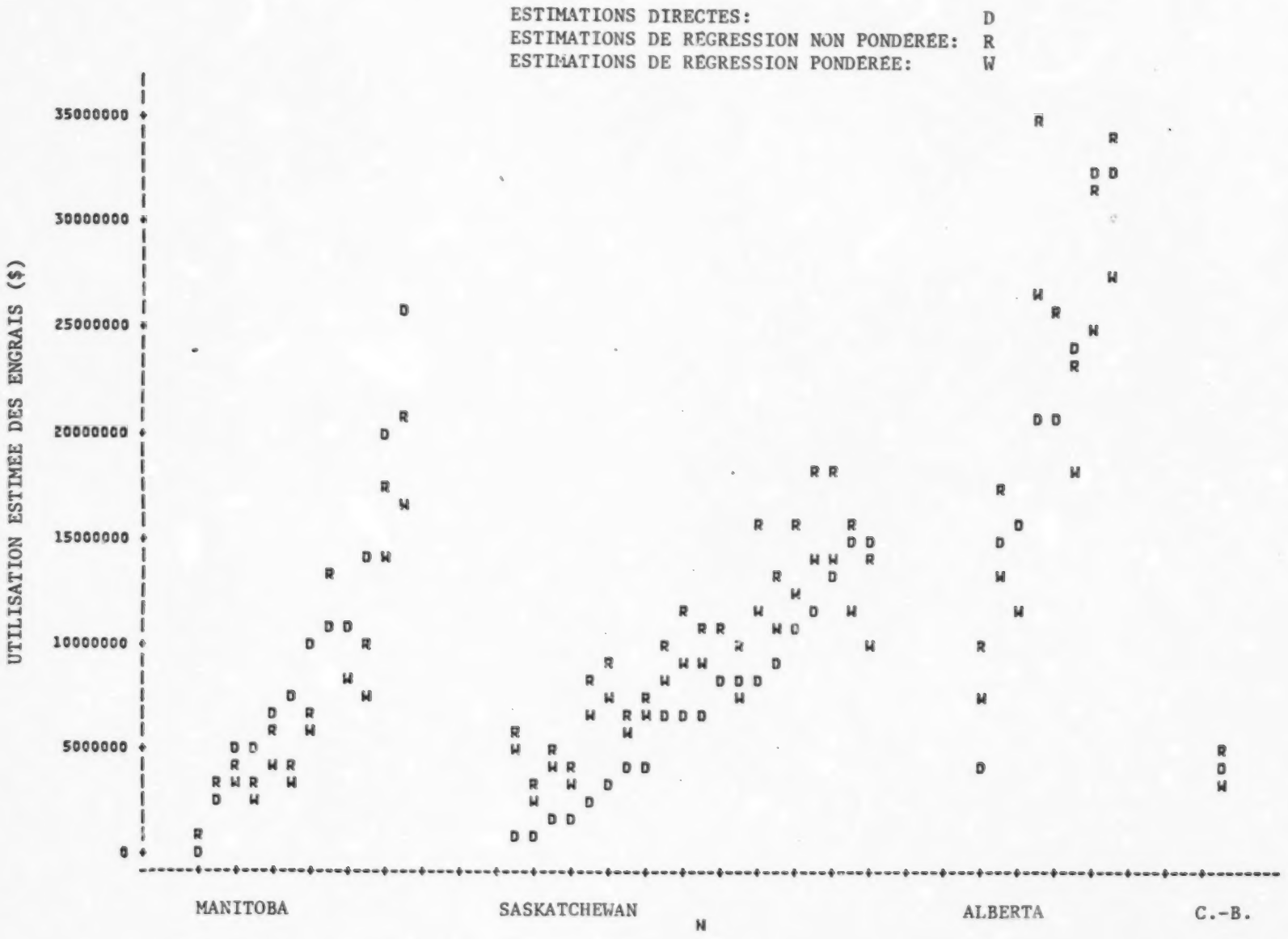
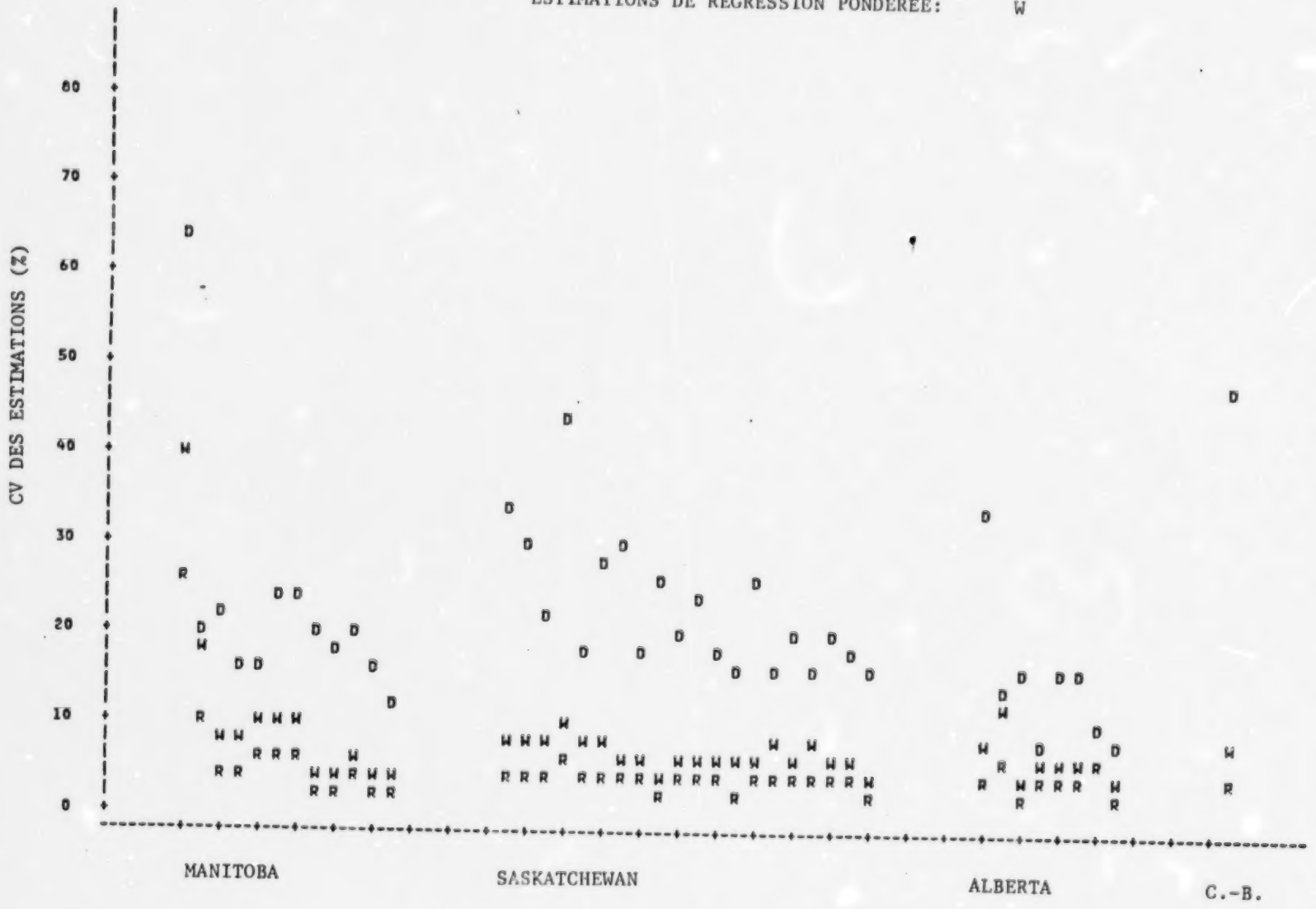


FIGURE 6: CV DES ESTIMATIONS DIRECTES, DES ESTIMATIONS DE RÉGRESSION NON PONDÉRÉE ET DES ESTIMATIONS DE RÉGRESSION PONDÉRÉE, PAR DISTRICT AGRICOLE



SECTION 2

La présente section ne constitue pas une analyse complète et structurée, mais fournit plutôt une description générale de certains aspects de l'utilisation de modèles linéaires pour l'inférence à partir de données d'enquêtes.

Dans son introduction, Nathan (1981) précise: "Quiconque veut entreprendre une analyse statistique doit tout d'abord définir les paramètres au sujet desquels il faut faire des inférences". La définition des paramètres dépend du modèle particulier que le statisticen choisit ou qu'on lui demande d'utiliser.

Dans la méthode traditionnelle d'analyse des enquêtes basée sur le plan d'échantillonnage, la population est considérée comme un ensemble de nombres fixes et non de variables aléatroires. C'est le plan d'échantillonnage qui introduit une composante aléatoire.

Les paramètres d'intérêt sont des paramètres de population finie, c'est-à-dire des fonctions de la population fixe. Ce point de vue me paraît analogue à celui d'un comptable qui s'intéresse aux recettes et aux dépenses réelles et non théoriques d'une entreprise.

Par contre, le modèle de superpopulation repose sur l'assertion selon laquelle la population est une réalisation d'une distribution de probabilité. Les paramètres peuvent dépendre de la distribution sous-jacente et non d'une réalisation particulière de cette distribution. Ce point de vue me paraît semblable à celui des dirigeants d'une entreprise qui peuvent s'intéresser à un modèle des recettes et des dépenses fondé sur une distribution de probabilité paramétrique. Ce n'est pas qu'ils croient que le modèle est vrai; au contraire, ils pensent que le modèle est faux. Ils cherchent plutôt à comprendre les processus qui déterminent les recettes et les dépenses.

La définition exacte des paramètres peut aider l'utilisateur à savoir s'il s'intéresse plus au résultat d'un processus ou qu'au processus lui-même.

Même quand une méthode basée sur le plan d'échantillonnage est utilisée, les paramètres de la population finie correspondent souvent à des estimateurs des paramètres de modèle de superpopulation. Il peut être utile d'estimer ces paramètres de la population finie, même quand la population n'est pas compatible avec le modèle correspondant de superpopulation. Toutefois, dans un tel cas, il faut s'assurer que les résultats de l'analyse sont interprétés seulement en fonction de la population et non comme des conclusions sur les mécanismes possibles qui produisent cette population.

L'éventail des modèles de superpopulation possibles est en fait sans limite. Un des modèles statistiques les plus fréquents est le modèle linéaire. Les paramètres de population finie qui correspondent à ce modèle, les proportions, les coefficients de corrélation et les coefficients de régression de la population, sont les outils généraux de l'inférence basée sur un plan d'échantillonnage. Dans les remarques suivantes sur les modèles linéaires, la notation s'inspire dans l'ensemble de celle de Nathan (1981).

Appelons la variables endogène Y et le vecteur de variables exogènes X (les variables auxiliaires). Notre choix d'un modèle linéaire pour exprimer la relation entre X et Y signifie que nous croyons que:

$$Y \triangleq X' \beta$$

pour un vecteur inconnu et selon notre interprétation du symbole "$\triangleq$".

Si par "$\triangleq$" nous entendons:

$$Y_i = X_i' \beta + e_i$$

où les e_i sont des variables gaussiennes indépendantes à moyenne nulle et à variance exprimée en fonction d'un multiplicateur constant (c'est-à-dire e_i $N(0, \sigma^2 k_i)$, k_i étant connu), nous avons alors le modèle classique de régression linéaire. Dans ce cas, l'inférence se fait par les moyens habituels et de façon relativement directe. Les poids d'échantillonnage ne seraient pas utilisés dans cette analyse et, si k_i est égal à une constante, la formule suivante produrait une estimation de β:

$$\hat{\beta} = (X_n' X_n)^{-1} X_n' Y_n$$

Par contre, si par "$\triangleq$" nous entendons quelque chose d'un peu différent, la voie à suivre pour l'inférence n'est pas aussi évidente. D'autres interprétations du modèle $Y \triangleq X'\beta$ pourraient être:

a. $Y_i = X_i' \beta + e_i$ où les e_i ne sont pas des variables gaussiennes, indépendantes, etc.

b. $Y_i = f(X_i)$ sans composante aléatoire, mais où $|f(X_i) - X_i'\beta|$ est, dans un certain sens, "petite" en moyenne.

c. $Y_i = f(X_i) + e_i$ où les e_i forment une composante aléatoire et $|f(X_i) - X_i'\beta|$ est "petite".

Le troisième modèle est assez général et inclut les deux autres modèles, ainsi que le modèle classique, comme cas spéciaux. Les hypothèses sur lesquelles ce modèle repose peuvent se résumer ainsi:

1) Il existe une relation fonctionnelle entre la variable endogène et les variables exogènes; on peut aussi ajouter une composante aléatoire à cette relation.

2) Il est possible de représenter cette relation fonctionnelle approximativement au moyen d'une relation linéaire.

3) Ce qui nous intéresse vraiment, c'est la "vraie" relation fonctionnelle, mais, faute de celle-ci, nous sommes prêts à faire des déductions à partir d'une approximation linéaire de cette relation.

Il faut bien réfléchir avant de décider quels paramètres nous voulons estimer et quel ensemble de paramètres est meilleur ou moins bon qu'un autre. La "meilleure" approximation linéaire d'une fonction dépendra tout probablement de la distribution des variables exogènes qui nous intéressent. (Par exemple, si nous voulons utiliser une droite comme approximation de la relation $f(X) = X^2$, la "meilleure" droite serait différente dans les cas où nous nous intéressons seulement aux valeurs positives de X ou exclusivement aux valeurs négatives de X.)

Dans un modèle de régression linéaire, la méthode classique pour comparer deux ensembles de paramètres est fondée sur la somme des carrés des résidus. Cette technique de comparaison est raisonnable dans le modèle classique où les erreurs sont gaussiennes, mais ce n'est pas nécessairement vrai dans d'autres situations. Toutefois, à cause de sa simplicité mathématique, cette solution peut sembler préférable.

Il arrive souvent que la distribuiton des variables exogènes à partir desquelles nous voulons produire une approximation de f(X) correspond à la distribution de ces variables dans la population échantillonnée et que, pour des raisons de simplicité mathématique autant que pour d'autres raisons, nous voulons utiliser la somme des carrés des résidus pour comparer les estimations des paramètres. Dans ce cas, l'estimateur pondéré:

$$\hat{\beta}_w = (X_n' W_n X_n)^{-1} X_n' W_n Y_n$$

corrige la représentation excessive dans l'échantillon des unités qui ont une probabilité de sélection élevée. Si le modèle classique est vrai et que $k_1 = 1$, la variance de $\hat{\beta}_w$ est:

$$\sigma^2 (X_n' W_n X_n)^{-1} X_n' W_n X_n (X_n' W_n X_n)^{-1} .$$

Ce résultat est un peut plus grand que la variance de $\hat{\beta}$, mais notons que, dans cette forme du modèle classique:

$$E(\hat{\beta}_w) = (X_n' W_n X_n)^{-1} X_n' W_n (X_n' \beta) = \beta$$

et $\hat{\beta}_w$ est donc également non biaisé.

En outre, $\hat{\beta}_w X_n$ est un estimateur convergent de Y_p basé sur le plan d'échantillonnage, alors que $\hat{\beta} X_n$ ne l'est pas. L'estimateur pondéré $\hat{\beta}_w$ semble donc présenter certains avantages par rapport à $\hat{\beta}$.

Pour faire des inférences basées sur le modèle, il nous faut encore une estimation de la variance résiduelle σ^2. L'espérance mathématique des carrés des résidus inclut un élément d'erreur et un élément de biais $(f(X) - X'\beta)$. Si le biais est relativement peu important en comparaison de l'erreur, il peut être raisonnable d'en faire abstraction et d'utiliser les carrés des résidus pour estimer σ^2. Faut-il alors utiliser les poids d'échantillonnage dans l'estimation de σ^2? Si on doit les utiliser, il faut d'une certaine manière les normaliser, peut-être de façon à ce que leur somme corresponde à la taille de l'échantillon (c'est-à-dire pour que le poids moyen soit égal à un). La variance résiduelle non pondérée (calculée à l'aide du diviseur approprié) semble également être une estimation raisonnable de σ^2. (En fait, une comparaison de l'estimation pondérée et de l'estimation non pondérée de la variance peut offrir un instrument diagnostique plus sensible pour vérifier la qualité du modèle linéaire à variance constante pour l'examen, proposé par Nathan (1981), de l'écart entre les estimations pondérée et non pondérée des coefficients, puisque le résultat de l'ajustement peut être essentiellement identique malgré les différences entre les coefficients.)

Il est implicite dans tout ce qui précède que les variables exogènes incluses dans le modèle ont déjà été choisies et qu'une décision a été prise au sujet du sous-ensemble d'unités d'enquête qui serviront à l'ajustement du modèle. Ni une ni l'autre de ces deux décisions ne peut se faire automatiquement. Les méthodes pas à pas classiques peuvent aider à déterminer le modèle qu'il faut utiliser, mais les statistiques F et les niveaux de signification sont inexacts si on emploie des progiciels généraux pour effectuer des régressions pondérées. Le choix du sous-ensemble d'unités d'enquête qui entreront dans l'ajustement du modèle doit supposer un compromis entre, d'une part, le besoin de minimiser le biais et, d'autre part, la nécessité d'obtenir des estimations précises de β à partir de toutes les données qu'on possède. Ainsi, ce sous-ensemble se composerait sans doute de toutes les unités d'enquête dans le domaine spécifié (ou de toutes les unités dans une classe d'unités d'enquête), mais peut-être également d'unités d'enquête dans d'autres domaines.

POOR

QUALITY

ORIGINAL